CLAIMS

1. A method for generating masks for data de-duplication from entity eponym data fields in a given set of data records , said data records each including an entity location data field, the method comprising:

for each data record, splitting each entity eponym into prefix-suffix combinations, and for each prefix, tallying matches with distinct entity locations, and tallying matches of distinct entity locations with a single derived suffix, and for each prefix and entity location combination, tallying distinct suffixes therefor; and

setting a threshold boundary wherein a prefix is defined as one of said masks when one or more tallies is indicative of different eponyms signifying a single one of said entities.

2. The method as set forth in claim 1, said setting a threshold boundary further comprising:

setting the threshold boundary wherein one or more tallying results is indicative of said entity eponym data field including variable data.

3. The method as set forth in claim 1, said setting a threshold boundary further comprising:

setting a threshold boundary wherein a tally of distinct suffixes is indicative of suffixes being information other than entity identity.

4. The method as set forth in claim 1, said setting a threshold boundary further comprising:

setting a threshold boundary where a ratio of a tally for said distinct suffixes to a tally for distinct entity locations is indicative of information other than entity identity.

5. The method as set forth in claim 1 further comprising:

applying an override function to said threshold boundary when a characteristic of said data record is indicative of a requirement for improving accuracy before a said prefix is defined as one of said masks.

6. The method as set forth in claim 1 further comprising:

prior to said splitting, creating a reduced data records sub-set by eliminating records having a unique entity eponym and entity location data pair.

7. The method as set forth in claim 1 further comprising:

generating a display showing each data record as each derived prefix and each related said entity location as a function of number unique suffixes concatenated with said each derived prefix as a function of number of each related said entity location.

8. The method as set forth in claim 1 wherein said de-duplication is a matching of each data record of a specific activity to a specific known entity of a plurality of known entities such that de-duplication of entities is minimized in a database of said plurality of known entities.

9. The method as set forth in claim 8 wherein said masks are generated as rules for ignoring variable data portions of a said entity eponym data field and assigning a respective data record therefor to said database based on non-variable data portions of said entity eponym data field.

10. The method as set forth in claim 9 further comprising:

maintaining said database by periodic application of said rules to a different said set of data records to be added to said database.

11. A method for partitioning a plurality of data packets in a database such that duplication of data groups is minimized, the method comprising:

selecting a primary identifier data field and a secondary identifier data field for each data packet;

for all data packets having a non-unique primary identifier data field, using heuristic procedures for splitting each primary identifier data into at least one prefix-suffix combination;

for each prefix, counting a first tally of how many distinct secondary identifier data fields occurs, and counting a second tally of how many distinct secondary identifier data fields occur with a single suffix, and for each prefix and each secondary identifier data field matched thereto, counting a third tally of how many distinct suffixes occur;

based on said first tally, said second tally and said third tally generating masks representative of prefixes applicable to said data packets having a non-unique primary identifier data field such that application of said masks assigns data packets having a non-unique primary identifier data field to associated common entities defined thereby; and

filing each of said data packets into a single file assigned to respective said associated common entities defined.

12. The method as set forth in claim 11 further comprising:

prior to said splitting, storing all data packets having a unique primary identifier data field in a respective unique data file, and

selecting one of a first subset of data packets wherein a primary identifier data field and a secondary identifier data field are substantially identical, and storing a remainder of said subset.

13. The method as set forth in claim 11 wherein said primary identifier data field is an intended unique entity name data field.

14. The method as set forth in claim 11 wherein said masks are generated to merge common entity name prefixes.

15. The method as set forth in claim 11 wherein said secondary identifier data field is a postal code data field.

16. The method as set forth in claim 11 further comprising:

retaining said masks as rules for cleaning dirty data portions of a data field by removing variable data segments therefrom.

17. A method of doing business comprising:

receiving a periodic log of transactions, each transaction being a data string including at least a name field and another identifier field;

selecting unique representative samples of said transactions;

for each of said samples, dissecting each name field into derived prefix and suffix combinations, and for each derived prefix and each prefix-another identifier combination, counting the number of distinct suffixes and storing a tally therefor; and

generating a mask from a specific prefix when the specific prefix meets a predefined decision criteria which is a function of said tally.

18. The method as in claim 17 wherein for each said derived prefix, counting prefix-another identifier combinations and storing a first tally therefor and counting prefix-distinct another identifier combinations and storing a second tally therefor, such that said predefined decision criteria is a function of said tallies.

19. A computer memory comprising:

for a given set of data records for a given set of entities, each of said data records having discrete data fields including an entity identification field and an entity location field,

computer code means for extracting a data pair from each of said records wherein said pair is defined as;

for each data pair, computer code means for splitting each entity identification data string into a plurality prefix-suffix combinations;

for each prefix, computer code means for tallying matches with a distinct entity location

data string, and computer code means for tallying matches of each distinct entity location data string with a single derived suffix;

for each prefix and entity location data string combination, computer code means for tallying distinct suffixes therefor;

computer code means for setting a threshold boundary wherein a prefix is defined as one of said masks when one or more tallies is indicative of a different entity identification data string signifying a single one of said entities; and

computer code means for applying said masks to said given set of data records such that each record is assigned to a single one of said given entities.

20, A system for data storage for a given set of segmented data records wherein each of said records includes a variable data segment and a fixed data segment, the system comprising:

means for splitting said variable data segment into sub-segments;

means for determining combinations of said sub-segments with said fixed data segment wherein said combinations are indicative of one of said sub-segments including a substantially constant data string portion and a variable data string portion;

means for comparing each constant data string portion to a boundary condition and for determining therefrom whether a constant data string portion is a substantially valid indicator for all said data records for assigning each of said data records to an associated data storage file created therefor;

means for applying each said valid indicator to said data records and for storing said data records in accordance therewith.

21. The system as set forth in claim 20 comprising:

means for concatenating each of said sub-segments with a correlated said fixed data segment and for forming a display representative of a distribution of all said combinations, and

means for setting a boundary condition between points of said distribution such that said points are divided between sub-segments forming a said substantially valid indicator and sub-segments not forming a said substantially valid indicator.